心理学报 2022, Vol. 54, No. 9, 1076−1092
Acta Psychologica Sinica

https://doi.org/10.3724/SP.J.1041.2022.01076

算法歧视比人类歧视引起更少道德惩罚欲*

许丽颖 [1] 喻 丰 [2] 彭凯平 [3]

([1] 清华大学马克思主义学院, 北京 100084) ([2] 武汉大学哲学学院心理学系, 武汉 430072)
([3] 清华大学社科学院心理学系, 北京 100084)

摘 要 算法歧视屡见不鲜, 人们对其有何反应值得关注。6 个递进实验比较了不同类型歧视情境下人们对算法歧视和人类歧视的道德惩罚欲, 并探讨其潜在机制和边界条件。结果发现: 相对于人类歧视, 人们对算法歧视的道德惩罚欲更少(实验 1~6), 潜在机制是人们认为算法(与人类相比)更缺乏自由意志(实验 2~4), 且个体拟人化倾向越强或者算法越拟人化, 人们对算法的道德惩罚欲越强(实验 5~6)。研究结果有助于更好地理解人们对算法歧视的反应, 并为算法犯错后的道德惩罚提供启示。

关键词 算法, 算法歧视, 道德惩罚, 自由意志信念, 拟人化

分类号 B849: C91

1 前言

歧视普遍存在, 性别、学历、种族、年龄之歧视常引发社会热议。面对歧视者, 大众倾向于道德谴责, 并希望其得到惩罚。在传统歧视事件中, 做出歧视的主体是人类。然而, 随着人工智能的发展和应用, 算法也成为了新的歧视主体。算法因其远超人类的计算能力和较低的成本而备受青睐, 并且已经逐步进入人类生活的各重要领域代替人类做出关键决策, 例如在医疗领域决定谁能够先得到器官捐赠(Freedman et al., 2020)、在金融领域决定投资者应该投资哪支基金(Harvey et al., 2017), 甚至在司法领域决定罪犯的风险等级并做出量刑(Hao, 2019)。不仅如此, 算法也因其能在一定程度上避免人类的主观性而被认为比人类决策者更准确公正(Grove et al., 2000)。然而, 算法虽然看似比人类更理性、更中立, 但算法决策也可能因训练数据集等问题而导致歧视(Borgesius, 2018)。如 Northpointe 公司所开发的对罪犯进行重复犯罪风险评估的 COMPAS 算法被发现存在种族歧视, 它会增加黑人被标记为累犯的可能性(Angwin et al., 2016)。谷歌的定向广告投放中也存在算法性别歧视, 与将用户的性别设置为男性相比, 将用户的性别设置为女性会导致高薪工作相关广告出现的次数更少(Datta et al., 2015), 类似的性别歧视也发生在科学、技术、工程和数学(STEM)领域的招聘广告投放算法中(Lambrecht & Tucker, 2019)。人们原本以为算法能够有助于减少乃至消除偏见, 但实际上相关的算法歧视案例不胜枚举, 在教育(Ferrero & Barujel, 2019)、医疗(Obermeyer et al., 2019)、消费(Angwin et al., 2015)等与人们生活息息相关的重要领域都有涉及。

面对人类歧视, 人们迫切地希望其受到道德惩罚, 那面对新型算法歧视, 人们是希望其受到同样惩罚吗? 为了回答这一问题, 本研究试图考察人们对人类歧视和算法歧视在道德惩罚欲上是否存在差异, 并且在此基础上进一步探讨造成其差异的潜在原因和边界条件。

1.1 人类歧视与算法歧视

歧视是指针对特定类别群体或其成员的无理负面行为(Al Ramiah et al., 2010), 这种行为不是由于群体及其成员应得或出于互惠, 而仅是由于其属

收稿日期: 2021-08-21
* 国家自然科学基金青年项目(72101132); 国家社科基金青年项目(20CZX059)。
通信作者: 喻丰, E-mail: psychpedia@whu.edu.cn

于特定类别(Correll et al., 2010)。与之类似, 算法歧视也与类别相关。当算法产生与受法律保护的类别变量(如种族和性别)相关的系统性差异时, 就被认为具有歧视行为(Bonezzi & Ostinelli, 2021), 比如亚马逊的招聘算法对女性简历评分更低(Dastin, 2018)。面对歧视这种有违公平、造成伤害的不道德行为(Haidt & Graham, 2007), 人们会产生道德反应, 即情感上的道德义愤(moral outrage; Batson et al., 2007)及行为倾向上的道德惩罚欲(desire for moral punishment; Hofmann et al., 2018)。道德惩罚是对不道德行为的制裁, 能够在一定程度上矫正已有的不道德行为并阻止未来的不道德行为, 因此在维持和加强道德体系上具有重要作用(Hofmann et al., 2018)。看到职场女性受上司歧视, 看到新冠疫情初期武汉人受到外地人歧视, 看到老年人在商店由于不会手机支付而受到店员歧视, 我们会愤怒, 意欲惩罚施加歧视者。

如果这些歧视行为的主体不是人类, 而是算法呢? 相对于人类歧视, 人们对算法歧视会产生较少的道德义愤, 这是由于人们会将较少的负面动机归因于算法(Bigman et al., 2020)。实际上, 如果从人类歧视和算法歧视所造成的后果来看, 算法歧视比人类歧视所造成的后果可能更为严重(Bigman et al., 2020)。以亚马逊公司的招聘为例, 可能有一定比例的人事经理会外显或内隐地歧视女性应聘者, 但个人影响毕竟有限; 若算法一旦应用, 则歧视所影响的应聘者可能成倍增长。因此如果单从后果上来看, 相对于人类歧视, 人们似乎可能会更想惩罚算法歧视。但本文试图从更本质的视角, 即人们对算法本身的知觉来探讨对算法歧视的反应。换句话说, 在歧视后果对等的情况下, 由于人们对算法和人类心智知觉的差异, 即认为算法相对于人类具有更少的自由意志, 因此人们对算法歧视的道德惩罚欲小于对人类歧视的道德惩罚欲。据此, 本文提出假设 1: 相较于人类歧视, 人们对算法歧视的道德惩罚欲更小。

1.2　自由意志信念与道德惩罚

面对不道德行为, 究竟是什么决定了我们是否想要对行为者进行道德惩罚? 当然, 自由意志(free will)是行为者对自身行为承担道德责任的必要条件(Nichols & Knobe, 2007)。简单来说, 自由意志就是自由行动的能力, 而自由行动意味着一个人可以在相同的情况下做出不同的选择和行为(Baumeister, 2014)。当一个人由于没有别的选择而只能做出不道德行为时, 对其进行谴责和惩罚显然是不合理的(Shariff et al., 2014), 因此道德惩罚会相应减少(Clark et al., 2014); 而如果人们要谴责和惩罚做出不道德行为的人, 则需要其至少有一定程度的自由意志。这也是为什么当违规者试图降低自身的负罪感并逃脱惩罚时, 常见的策略就是将其行为描述为无能为力且无法避免的选择(Baumeister et al., 1990)。心理学家不甚关心自由意志是否真实存在, 反而更关心人们是否相信自由意志存在, 即自由意志信念(belief in free will) (Baumeister, 2008)。自由意志信念薄弱会造成负面后果, 例如减少助人行为且攻击性增强(Baumeister et al., 2009)、增加欺骗行为(Vohs & Schooler, 2008)以及降低自控力(Rigoni et al., 2012)等。更重要的是, 削弱人们的自由意志信念或者提供关于违规者缺乏自由行动能力的相关证据, 会影响人们对道德责任信念的推脱, 进而导致人们做出更多不道德的行为(Shariff et al., 2014), 也会导致人们对违规者道德惩罚的减少(如 Aspinwall et al., 2012)。

实际上, 大多数人都相信人类有自由意志(Nahmias et al., 2005), 因此当歧视的行为者是人类时, 人们更有可能认为歧视行为是出于其自由意志的结果, 从而产生较强的道德惩罚欲。那算法呢? 虽然现阶段的算法缺乏完全的自由意志和自主性, 但与“客观”自主性(即人工智能是否有自主性)相比, “主观”自主性(即人们是否认为人工智能具有自主性)对于其道德责任的影响似乎更加重要(Wegner & Gray, 2017)。因此, 歧视作为一种不道德行为会引发道德惩罚欲, 而道德惩罚欲的大小则受到人们认为歧视者在多大程度上拥有自由意志的影响。现有研究表明, 人们对人类和算法等人工智能的心智知觉(mind perception)不同。与人类相比, 人工智能具有中等程度的能动性(agency; 即自主计划行动等心理能力), 并且具有较低程度的体验性(experience; 即能够体验情绪等心理能力) (Bigman & Gray, 2018; Gray et al., 2007)。也就是说, 算法虽然具有一定程度的自主行为能力, 但并不会被视为与人类具有同等程度的自由意志(Weisman et al., 2017; Shariff et al., 2014)。总之, 与人类相比, 人们认为算法具有较少的自由意志。基于上文所述, 本文认为人们对算法歧视比对人类歧视有更少的道德惩罚欲, 这是由于认为算法比人类具有更少的自由意志。据此, 本文的假设 2 为: 自由意志信念在歧视主体(人类 vs. 算法)对道德惩罚欲的影响中起

中介作用。

当然，需要说明的是，本文所提出的自由意志信念并非人们对不同歧视主体(人类 vs. 算法)产生不同道德惩罚欲的唯一解释机制。例如，算法没有人一样的不良动机(Bigman et al., 2020)、算法本身应承担的责任较小(多数责任可归于编制算法者)、算法受惩罚并不能促使其进步等，这些都能够在一定程度上作为解释机制。但本文之所以重点关注自由意志，主要是由于上述可能的解释机制均与自由意志密切相关。第一，具有自由意志的个体作出不道德行为可能说明其具有不道德的动机(该个体是“坏”的)，即判断个体具有自由意志可能是我们推测其动机的必要条件(如 Laming, 2004)。第二，具有自由意志的个体能够自主选择，也应当自主承担责任，即自由意志是个体承担责任的必要条件(如 Sinnott-Armstrong, 2014)。第三，具有自由意志的个体可能能够理解惩罚并反思，对其不道德行为进行惩罚更可能促使其产生积极变化，因此或许个体在一定程度上具有自由意志亦是惩罚能够对个体产生积极促进作用的必要条件。于是，我们认为自由意志与动机、责任和惩罚效果等因素密切相关，并且自由意志信念的解释机制在某种意义上可能更为基础，能够在一定程度上涵盖以上所述的其它解释机制。因此，本文对不同歧视主体(人类 vs. 算法)如何影响道德惩罚欲的机制探讨将着重检验自由意志信念的作用。

此外，可能还存在与自由意志无关的竞争假设。首先，人的行为更容易被解释，而算法复杂和不透明。被试因为算法的内在逻辑难以甄别，也就更难判定歧视行为是不道德的，甚至可能认为其具有合理性，进而对其表现出宽容。其次，人无法真正惩罚算法，即惩罚算法是不切实际的，但人可以惩罚人类。换言之，所谓惩罚算法是惩罚算法的载体，并非惩罚算法本身。鉴于很难真正惩罚算法，因此当算法出现歧视后，人不怎么愿意惩罚它。鉴于以上与自由意志无关的竞争假设存在，我们将会通过 4 个调节效应研究(研究 3~6)将之进行排除。

1.3 拟人化

自由意志通常情况下被当作人的特征(Waytz et al., 2010)，讨论算法是否具有自由意志即是在将算法拟人化。拟人化(anthropomorphism)是指“将人类特征、动机、意向或心理状态赋予非人对象”(Epley et al., 2007)。人工智能尤其是算法对于人们来说相对抽象，在很多情况下设计者会以拟人化的方式将其呈现，人们也倾向于以拟人化的方式对其进行知觉。人工智能的拟人化能够在一定程度上增进信任(Waytz, Heafner, & Epley, 2014)，但过度拟人化也可能会诱发恐怖谷(uncanny valley)效应(Mori, 1970)，使积极态度急转直下。人们会根据人工智能的外表来推断其心智，人工智能越像人，人们就越倾向于认为其有类人的心智(Bigman et al., 2019)。当然，拟人化人工智能的方式本身也包括赋予其人类的心理状态如自由意志等，使人工智能的行为看起来是自由选择的结果。不过，人们的拟人化倾向也是存在个体差异的(Waytz et al., 2010)，面对同样一个人工智能，有的人会更倾向于将其拟人化，而有的人则不会。拟人化倾向的个体差异影响广泛(Epley & Waytz, 2010)，越多将算法拟人化，则越可能认为算法具有某种程度的自由意志或者自主性，并让人们可以对其进行道德归责(Gray et al., 2007)。

总之，人们越倾向于拟人化人工智能，就会认为其拥有更多的自由意志，从而需要为其行为承担更多的道德责任乃至惩罚(Bigman et al., 2019; Waytz, Cacioppo, & Epley, 2014)。因此，拟人化倾向的个体差异以及算法本身的拟人化程度都会影响到人们是否更加以拟人的方式看待算法，同时也影响到人们对于道德违规算法的道德惩罚欲。据此，本文提出假设 3：拟人化在歧视主体(人类 vs. 算法)对道德惩罚欲的影响中起调节作用。具体而言，在个体的拟人化倾向方面，对于拟人化倾向较低的人而言，他们对人类歧视的道德惩罚欲大于对算法歧视的道德惩罚欲，对于拟人化倾向较高的人来说，由于将算法更多当作人来看待，因而对人类歧视和算法歧视的惩罚欲没有显著差异；而在算法本身的拟人化方面，算法的拟人化程度越高，人们对算法歧视与对人类歧视的道德惩罚欲差异越小。

1.4 研究概览

综上所述，本研究旨在考察人们对人类歧视和算法歧视的道德惩罚欲是否存在差异，并在此基础上进一步探讨其心理机制和边界条件。本研究的基本假设是：人们对算法歧视的道德惩罚欲小于对人类歧视的道德惩罚欲，这一效应受到自由意志信念的中介和拟人化的调节。本研究采用递进的 6 个实验来验证假设。6 个实验均采用情境实验的方式，为被试呈现人类或算法的歧视行为并测量其道德惩罚欲，涉及到的歧视包括性别歧视(实验 1 和实验 6)、学历歧视(实验 2)、民族歧视(实验 3 和实验

4)和年龄歧视(实验 5), 并涵盖了具有代表性的全国范围被试和大学生被试。具体而言：实验 1 验证人们对算法歧视的道德惩罚欲是否小于对人类歧视的道德惩罚欲; 实验 2 探究其中的心理机制, 验证自由意志信念在歧视主体(人类 vs. 算法)影响道德惩罚欲中的中介作用; 实验 3 通过操纵被试的自由意志信念, 进一步检验自由意志信念是否是导致人们产生不同道德惩罚欲的机制; 实验 4 则通过具体操纵人们对算法的自由意志信念, 再一次检验自由意志信念是否是导致人们对不同歧视主体(人类 vs. 算法)有不同道德惩罚欲的机制; 实验 5 探索可能的边界条件, 验证拟人化倾向在歧视主体影响道德惩罚欲中的调节效应; 实验 6 则直接对算法的拟人化程度进行操纵, 进一步验证拟人化在歧视主体影响道德惩罚欲中的调节效应。

2　实验 1：算法歧视引发更少道德惩罚欲

实验 1 的目的是初步探讨与人类歧视相比, 算法歧视是否会引发人们更少的道德惩罚欲。我们采用网络情境实验的方法, 将被试随机分配至人类组和算法组, 分别阅读人类歧视和算法歧视的情境材料并报告道德惩罚欲, 以此比较人们对人类歧视和算法歧视的道德惩罚欲差异。

2.1　方法

2.1.1　被试

本研究首先使用 G*Power 3.1 软件(Faul et al., 2007)计算研究所需样本量。以独立样本 t 检验为统计方式, 显著性水平 $\alpha = 0.05$ 且中等效应量($d = 0.5$)时, 为了达到 90%统计检验力, 本实验至少需要 172 名被试。通过在 Credamo 平台发布实验, 实时剔除没有通过注意检查的被试数据并滚动采集, 最终得到 172 份有效数据, 包括男性 76 名(44.2%), 女性 96 名(55.8%), 平均年龄 M = 28.33 岁, SD = 4.26 岁。参与实验的被试被随机分派到人类组和算法组, 其中人类组 85 人, 算法组 87 人。所有被试均自愿参加实验且知情同意, 通过注意检查的被试在实验结束之后获得相应实验报酬。

2.1.2　实验设计与程序

实验 1 为单因素两水平被试间实验设计, 两组分别为人类组和算法组, 所有被试被随机分配到其中一组。首先, 所有被试均阅读性别歧视的情境材料(下划线内容为人类组材料, 括号中为算法组材料)：“李亮和何萍夫妻二人都申请了同一银行的信用卡, 夫妻双方都对自己的财产拥有平等的所有权, 并且收入相同。银行审理人(算法)对二人的申请进行评估, 最终给予李亮五万元的额度, 而何萍却只有三万元额度”, 情境材料改编自 Bigman 等人(2020)的研究。为了确保被试认真阅读并理解了情境材料的内容, 被试在阅读完情境材料后被要求回答注意检查题目(例如“对李亮和何萍进行信用卡额度评估的是？”1 = 银行审理人, 2 = 算法), 如未正确回答该题目, 该被试将在 Credamo 数据平台上被拒绝, 平台将递补搜集其他被试以满足样本量需求。

在阅读完情境材料并进行注意检查后, 被试填写了道德惩罚欲问卷。我们采用 Hofmann 等人(2018)对道德惩罚欲的测量, 请被试回答以下 3 个题目(括号中为算法组题目)：“你认为这个银行审理人(算法)应该为这种行为受到多大程度的道德惩罚？”、“你在多大程度上想要去惩罚这个银行审理人(算法)？”、“你在多大程度上认为应该要求这个银行审理人(算法)恢复因其不道德行为所造成的损害？”3 个题目均采用李克特 7 点量表计分(从“1 = 一点也不”到“7 = 非常”), 得分越高表明被试对情境中的人类(或算法)的道德惩罚欲越强。实验 1 中该测量的内部一致性信度 Cronbach $\alpha = 0.87$。

考虑到不同被试对算法的看法和知识可能存在差异, 从而影响到其对算法歧视的道德惩罚欲, 因此为了排除相关可能的影响因素, 被试还被要求报告他们对算法的熟悉程度(“你对算法有多熟悉？”, 从“1 = 一点也不熟悉”到“5 = 非常熟悉”)、了解程度(“与普通中国人相比, 你认为你对算法有多了解？”, 从“1 = 一点也不了解”到“5 = 非常了解”)和喜爱程度(“你有多喜欢算法？”, 从“1 = 一点也不喜欢”到“5 = 非常喜欢”)。其中熟悉程度和了解程度的题项改编自 Leo 和 Huh (2020)的研究, 喜爱程度的题项改编自 Godspeed 量表中的条目(Bartneck et al., 2009)。最后, 被试报告了性别和年龄两项人口统计学信息。

2.2　结果

独立样本 t 检验结果显示, 人类组的道德惩罚欲评分(M = 5.29, SD = 0.99)高于算法组(M = 4.97, SD = 1.34), 差异呈边缘显著, $t(170) = 1.82$, $p = 0.073$, Cohen’s $d = 0.27$。为了验证结果的稳健性, 将被试的性别(男 = 1, 女 = 2)和年龄作为协变量进行控制, 方差分析结果显示, 人类组的道德惩罚欲评分仍然高于算法组, 差异呈边缘显著, $F(1, 168) = 3.22$, $p = 0.075$, $\eta_p^2 = 0.019$。为了进一步排除年龄和

性别可能对结果的影响，我们分别对其进行了相关分析和独立样本 t 检验，结果发现年龄与道德惩罚欲评分相关不显著(r = 0.01, p = 0.853)，男性和女性的道德惩罚欲无显著差异，t(170) = 0.83, p = 0.408。

为了排除被试对算法的熟悉程度、了解程度和喜爱程度可能对结果的影响，我们将算法组的道德惩罚欲评分与这些变量进行相关分析，结果表明相关均不显著，$r_{熟悉}$ = −0.13, $r_{了解}$ = −0.10, $r_{喜爱}$ = −0.15, ps > 0.05。

2.3 讨论

实验 1 初步验证了算法歧视相比于人类歧视会引发人们更少的道德惩罚欲，并且排除了被试对算法的熟悉程度、了解程度和喜爱程度的可能影响。但实验 1 只涉及到一种歧视类型即性别歧视，并且未探索其中深层的心理机制。鉴于此，实验 2 的情境设置于算法歧视常见的招聘领域，重点考察其中可能存在的学历歧视现象，拟在进一步检验实验 1 结果稳健性的基础上试图发现自由意志信念的中介作用。

3 实验 2：自由意志信念的中介作用

实验 2 在实验 1 的基础上丰富了歧视的类型，加入了对学历歧视的考察，并且进一步探讨现象背后的深层机制，验证自由意志信念可能在其中扮演的中介作用。

3.1 方法

3.1.1 被试

对于本实验适用的独立样本 t 检验，取中等效应量 d = 0.5，显著性水平 α = 0.05，通过 G*Power 3.1 软件(Faul et al., 2007)计算本实验所需样本量结果表明，172 名被试才能达到 90%统计检验力。通过 Credamo 平台招募被试，实时剔除没有通过注意检查的被试数据并滚动采集，最终得到 172 份有效数据。这 172 名被试的平均年龄为 28.14 ± 6.21 岁，其中女性 104 名，占比为 60.5%，男性 68 名，占比为 39.5%。被试被随机分派到人类组(86 人)和算法组(86 人)。所有被试在实验开始之前均仔细阅读了实验说明并知情同意，有效数据被试在实验结束后获得实验报酬。

3.1.2 实验设计与程序

同实验 1，实验 2 也为单因素两水平被试间实验设计。被试首先阅读学历歧视的情境材料(下划线内容为人类组材料，括号中为算法组材料)：“在去年的秋季招聘中，韦蓝公司的人力资源经理李原负责(使用算法)进行招聘。招聘结束后，公司发现李原(算法)对应聘者的简历进行筛选时存在学历偏差，筛掉了所有硕士研究生以下学历的应聘者，而公司的大部分岗位对学历并无硬性要求。这阻碍了许多有才华、有能力、但不具有研究生学历的人获得该公司的工作”，情境材料改编自 Bigman 等人(2020)的研究。

在阅读完情境材料并进行注意检查后，两组被试分别报告了对人类歧视或算法歧视的道德惩罚欲，测量条目同实验 1 (Hofmann et al., 2018)。在实验 2 中，道德惩罚欲 3 个项目的内部一致性信度 Cronbach α = 0.87。然后，我们测量了被试对于情境中做出学历歧视行为的人类或算法的自由意志信念。采用改编后的自由意志量表(free will inventory; Nadelhoffer et al., 2014)，共 5 个条目(α = 0.86)，例如“李原(算法)是有自由意志的”，均为李克特 7 点计分(1 = 强烈反对, 7 = 强烈同意)，得分越高表明被试认为情境中的人类(或算法)有更多自由意志。最后，被试报告了性别、年龄和受教育程度三项人口统计学信息。

3.2 结果

3.2.1 歧视行为主体对道德惩罚欲的影响

独立样本 t 检验结果显示，人类组的道德惩罚欲评分(M = 5.11, SD = 1.14)显著高于算法组(M = 4.60, SD = 1.54), t(170) = 2.44, p = 0.016, Cohen's d = 0.38。

为了进一步验证结果的稳健性，将被试的性别(男 = 1，女 = 2)、年龄和教育程度(小学及以下= 1，初中 = 2，普高/中专/技校/职高 = 3，专科 = 4，本科 = 5，硕士研究生 = 6，博士研究生 = 7)作为控制变量，方差分析结果显示，人类组的道德惩罚欲评分仍然显著高于算法组，F(1, 167) = 5.96, p = 0.016, η_p^2 = 0.03。

3.2.2 自由意志信念的中介效应

为了探索歧视主体对道德惩罚欲影响的心理机制，我们使用 Hayes (2013)提供的 SPSS 插件 PROCESS (Model 4)，以歧视主体为自变量(人类组 = 0，算法组 = 1)，自由意志信念为中介变量，道德惩罚欲为因变量，设定 Bootstrap 样本量为 5000，采用偏差校正的方法，选取 95%置信区间进行中介效应检验。数据结果显示，自由意志信念的中介效应值为−0.56，95%的 Bootstrap 置信区间为[−0.95, −0.21]，不包含 0，表明中介作用显著；并且在控制

中介变量后, 歧视主体对道德惩罚欲的直接效应为 0.06, 95%的 Bootstrap 置信区间为[−0.44, 0.55], 包含 0, 表明其直接效应不再显著, 自由意志信念在歧视主体对道德惩罚欲的影响中起完全中介作用。为了进一步验证中介效应的稳健性, 我们又使用传统逐步回归方法进行了中介效应分析(温忠麟 等, 2004), 结果见图 1。

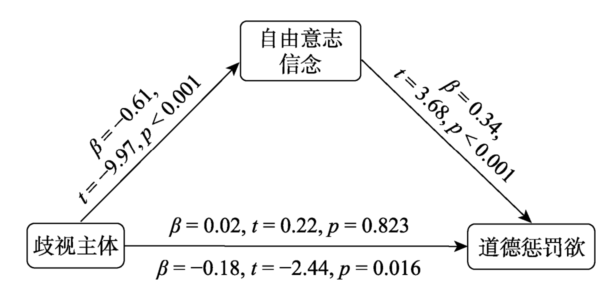


图 1　自由意志信念的中介作用

3.3　讨论

与实验 1 结果相一致, 实验 2 再次验证了算法歧视相比于人类歧视会引发人们更少的道德惩罚欲, 并且进一步发现了自由意志信念在其中的中介作用, 即人们认为算法相比人类有更少的自由意志, 因此不倾向于对其进行道德惩罚。因此, 实验 1 和实验 2 对于本文的主要假设, 即算法歧视比人类歧视引发更少道德惩罚欲, 提供了稳定一致的支持, 并且对自由意志信念的中介效应进行了初步验证。为进一步检验实验 1 和实验 2 结果的稳健性, 实验 3 将实验情境材料设置为民族歧视情境。此外, 为了进一步验证其中的心理机制, 即自由意志信念是造成不同道德惩罚欲的原因, 拟在实验 3 中操纵被试的自由意志信念。我们预测, 若启动被试“不存在自由意志”的信念, 那么歧视主体对道德惩罚欲的影响将会消失。

4　实验 3: 操纵自由意志信念

为了增加实验结果的稳健性, 实验 3 再次丰富了歧视类型, 关注民族歧视问题, 并且通过对被试自由意志信念的操纵, 进一步探讨自由意志信念是否是造成道德惩罚欲差异的机制。

4.1　方法

4.1.1　被试

采用 G*Power 3.1 软件(Faul et al., 2007)计算本实验所需样本量, 对于本实验适用的双因素方差分析, 取中等效应量 f = 0.25, 显著性水平 α = 0.05, 组数为 4, 要达到 85%的统计检验力至少需要 201 名被试。考虑到可能会出现少量不认真填答或未通过注意检查的无效数据, 我们共招募了 231 名来自某高校的本科生参加实验, 完成实验可获得相应学分。实验通过 Qualtrics 平台开展, 被试在实验开始前均详细阅读了实验说明并知情同意, 有 26 名被试的回答不符合要求或未通过注意检查, 最终得到 205 份有效数据。这些有效被试的平均年龄为 19.18 岁(SD = 0.81 岁), 女性 77 名(占 37.6%)。

4.1.2　实验设计与程序

实验 3 为 2 (歧视行为主体: 人类 vs.算法) × 2 (自由意志信念: 高 vs.低)被试间实验设计, 所有被试被随机分配到四个组的其中一组。

首先, 被试阅读自由意志信念的操纵短文。在低自由意志信念组, 被试阅读到一篇如下题为“科学表明自由意志并不存在”的短文, 为了增加短文的真实性和可信性, 我们标注了短文作者是一位名为“克里斯·惠灵顿”的博士。短文主要阐述了科学表明人们的所作所为都是他们大脑中简单物理过程的产物, 自由意志只是一种幻觉。在高自由意志信念组, 被试则阅读到如下一篇题为“科学表明存在自由意志”的短文, 短文同样标注了作者。短文主要阐述了科学表明人们的所作所为大多是他们做出的决定和自由意志的产物, 自由意志不是一种幻觉。

在阅读完短文之后, 所有被试被要求对该短文写一个简短的总结, 不少于 50 字。自由意志信念的操纵翻译改编自 Mackenzie 等人(2014)的研究。为了检查自由意志信念的操纵是否有效, 在写完简短总结后, 被试被要求回答“你在多大程度上相信存在自由意志? ” (1 = 一点也不相信, 9 = 完全相信)。

然后, 被试阅读民族歧视的情境材料(下划线内容为人类组材料, 括号中为算法组材料): “在去年的秋季招聘中, 韦蓝公司的人力资源经理张沛负责(使用算法)进行招聘。招聘结束后, 公司发现张沛(算法)对应聘者的简历进行筛选时存在民族偏差, 筛掉了所有少数民族的应聘者, 留下的都是汉族人, 而公司的所有岗位对民族并无要求。这阻碍了许多有才华、有能力的少数民族应聘者获得该公司的工作”, 情境材料改编自 Bigman 等人(2020)的研究。

接着, 在阅读完情境材料并进行注意检查后, 两组被试分别报告了对人类歧视或算法歧视的道德惩罚欲, 测量条目同实验 1 (Hofmann et al., 2018), 在实验 3 中, 道德惩罚欲测量的内部一致性信度 Cronbach α = 0.86。最后, 被试报告了性别、年龄和民族三项人口统计学信息。

4.2 结果

4.2.1 操纵检查

独立样本 t 检验结果显示, 低自由意志信念组的自由意志信念(M = 5.85, SD = 1.90)显著低于高自由意志信念组(M = 6.54, SD = 1.49), $t(203)$ = −2.88, p = 0.004, Cohen's d = −0.40。说明自由意志信念操纵有效。

4.2.2 歧视行为主体和自由意志信念对道德惩罚欲的交互作用

以歧视行为主体(人类组 = 0, 算法组 = 1)和自由意志信念(低自由意志信念组 = 0, 高自由意志信念组 = 1)作为自变量, 以道德惩罚欲作为因变量进行方差分析。数据结果表明, 人类组的道德惩罚欲评分(M = 4.59, SD = 1.46, 95% CI [4.31, 4.87])显著高于算法组(M = 4.17, SD = 1.51, 95% CI [3.87, 4.46]), $F(1, 201)$ = 4.01, p = 0.047, η_p^2 = 0.02), 高自由意志信念组的道德惩罚欲评分(M = 4.61, SD = 1.26, 95% CI [4.36, 4.86])显著高于低自由意志组(M = 4.17, SD = 1.67, 95% CI [3.85, 4.49]), 差异呈边缘显著, $F(1, 201)$ = 3.83, p = 0.052, η_p^2 = 0.02), 歧视行为主体和自由意志信念的交互作用显著, $F(1, 201)$ = 4.57, p = 0.034, η_p^2 = 0.02。简单效应分析发现, 在高自由意志信念组, 算法组的道德惩罚欲评分(M = 4.14, SD = 1.46, 95% CI [3.71, 4.58])显著低于人类组(M = 4.99, SD = 0.91, 95% CI [4.60, 5.39]), $F(1, 201)$ = 8.19, p = 0.005, η_p^2 = 0.04; 在低自由意志信念组, 算法组与人类组的道德惩罚欲评分无显著差异, $F(1, 201)$ = 0.01, p = 0.922, η_p^2 < 0.001 (见图 2)。

在将被试的性别(男 = 1, 女 = 2)、年龄和民族(汉族 = 1, 少数民族 = 2)作为协变量进行控制之后, 人类组的道德惩罚欲评分仍然显著高于算法组, $F(1, 198)$ = 4.52, p = 0.035, η_p^2 = 0.02), 高自由意志信念组的道德惩罚欲评分仍然显著高于低自由意志信念组, $F(1, 198)$ = 4.89, p = 0.028, η_p^2 = 0.02), 歧视行为主体和自由意志信念的交互作用仍然显著, $F(1, 198)$ = 4.88, p = 0.028, η_p^2 = 0.02。

4.3 讨论

实验 3 通过对被试自由意志信念的操纵, 进一步验证了自由意志信念是造成道德惩罚欲差异的机制, 发现只有在自由意志信念较高时, 不同歧视行为主体(人类 vs.算法)才会引发不同程度的道德惩罚欲; 而当自由意志信念较弱时, 人们对不同歧视主体的道德惩罚欲差异不显著。但实验 3 存在一些不足之处, 首先, 实验 3 未能充分检验本研究所提出的机制, 即"人们对算法的道德惩罚欲高于对人类的道德惩罚欲是由于人们认为算法相比人类缺乏自由意志"; 其次, 实验 3 对自由意志信念的操纵并未影响被试对算法的惩罚欲, 因此这种操纵可能并未对人们关于算法的自由意志信念产生影响。因此, 为了更直接地检验对算法的自由意志信念是否是造成人们对不同歧视主体(人类 vs. 算法)有不同道德惩罚欲的机制, 实验 4 将直接对人们关于算法的自由意志信念进行操纵。

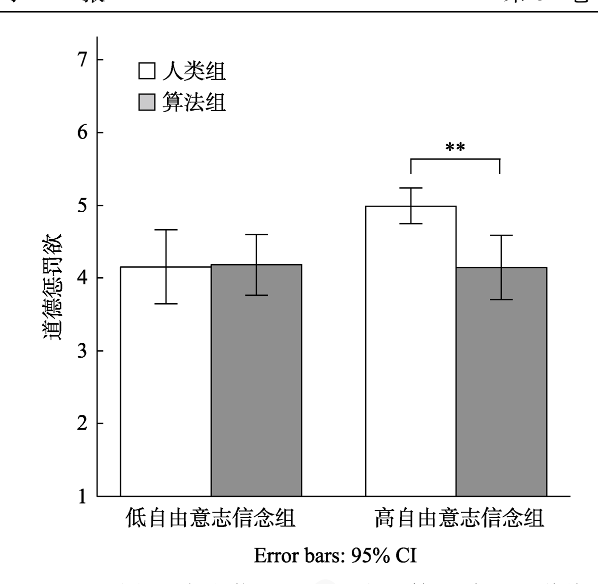


图 2 不同自由意志信念组对人类和算法歧视的道德惩罚欲评分

注: $^*p < 0.05$, $^{**}p < 0.01$.

5 实验 4: 操纵被试对算法的自由意志信念

实验 4 通过实验操纵的方式提升被试对算法的自由意志信念, 即采用单因素三水平设计(人类/算法/相信算法有自由意志), 考察人类组与相信算法有自由意志组之间的差异是否较人类组与算法组的差异有所减小。

5.1 方法

5.1.1 被试

采用 G*Power 3.1 软件(Faul et al., 2007)计算本实验所需样本量, 对于本实验适用的单因素三水平方差分析, 取中等效应量 f = 0.25, 显著性水平 α = 0.05, 组数为 3, 要达到 90%的统计检验力至少需要 207 名被试。考虑到可能会出现少量未完成或未通过注意检查的无效数据, 我们共招募了 247 名来自两所高校的本科生参加实验, 完成实验可获得相应学分。实验通过 Qualtrics 平台开展, 被试在实验

开始前均详细阅读了实验说明并知情同意，有 37 名被试未完成实验或未通过注意检查，最终得到 210 份有效数据。这些有效被试的平均年龄为 19.12 岁(SD = 1.28 岁)，女性 106 名(占 50.5%)。

5.1.2　实验设计与程序

实验 4 为单因素三水平被试间实验设计，两组分别为人类组、相信算法有自由意志组以及算法组，所有被试被随机分配到其中一组。

首先，被试阅读民族歧视的情境材料，实验 4 仍然使用实验 3 中所用的民族歧视情境，但具体不同之处有三点。第一，为了更好地排除被试对算法的理解程度可能对实验结果造成的影响，两个算法组在民族歧视情境描述之前均加入了对“算法”含义的解释及举例说明(改编自维基百科和 Merriam-Webster)，以确保被试在完成实验前理解算法的含义；第二，实验 4 中对人类的表述有所不同，由于人名(如“张沛”)属于具体明确的对象，而“算法”则属于宽泛的概念，无具体明确的对象，因此为了统一人类和算法情境表述的具体/抽象程度，实验4中的人类组的歧视主体仅表述为“人力资源经理”；第三，作为对被试是否相信算法有自由意志的操纵，相信算法有自由意志组的被试还阅读了以下关于情境中公司招聘所用算法的介绍：

韦蓝公司的算法接受过训练，可以根据应聘者的个人情况进行简历筛选。该算法的独特之处在于其是一个有自由意志的算法。也就是说，该算法的决策完全由其自己做出，并且其有能力做出不同的选择。

该操纵改编自 Kim 和 Duhachek (2020)对于人工智能是否有意识的操纵。为了检验对算法自由意志信念操纵的有效性，被试需要对情境中算法的自由意志进行评分(“你认为该算法在多大程度上拥有自由意志？”，从“1 = 一点也没有”到“7 = 非常多”)。

在阅读完情境材料并进行操纵和注意检查后，三组被试分别报告了对人类或算法的道德惩罚欲。需要提到的是，为了改进前面 3 个实验的道德惩罚欲测量条目中“道德惩罚”、“不道德行为”等表述可能对被试作答的影响，本实验对测量条目进行了相应修改，将第一道题和第三道题分别修改为：“你认为该人力资源经理(算法)应该为这种行为受到多大程度的惩罚？”、“你在多大程度上认为应该要求该人力资源经理(算法)恢复因其行为所造成的损害？”，评分和计分方式同实验 1 (Hofmann et al., 2018), α = 0.82。最后，被试报告了性别、年龄和民族三项人口统计学信息。

5.2　结果

5.2.1　操纵检查

独立样本 t 检验结果显示，相信算法有自由意志组对算法的自由意志信念(M = 3.70, SD = 1.73)显著高于算法组(M = 2.65, SD = 1.34), $t(137) = 4.00$, $p < 0.001$, Cohen’s $d = 0.68$。说明我们对被试关于算法的自由意志信念操纵有效。

5.2.2　关于算法的自由意志信念的影响

以组别(人类组 = 1，相信算法有自由意志组 = 2，算法组 = 3)作为自变量，道德惩罚欲作为因变量进行方差分析，结果显示组别的主效应显著，$F(2, 207) = 9.03$, $p < 0.001$, $\eta_p^2 = 0.08$。计划对比(planned contrast)分析表明，算法组的道德惩罚欲(M = 3.94, SD = 1.45)显著低于相信算法有自由意志组(M = 4.56, SD = 1.62)和人类组(M = 4.98, SD = 1.35), $ps < 0.05$，但人类组与相信算法有自由意志组的道德惩罚欲无显著差异，$p = 0.10$ (如图 3)。

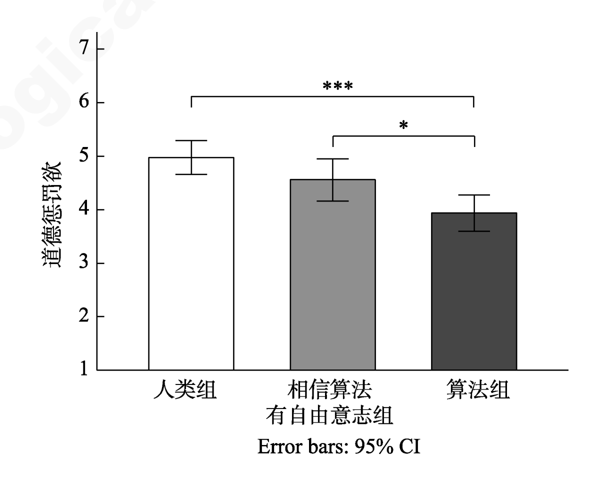


图 3　不同歧视主体组的道德惩罚欲评分

注：$^{*}p < 0.05$, $^{***}p < 0.001$.

以组别作为自变量，性别(男 = 1，女 = 2)、年龄和民族(汉族 = 1，少数民族 = 2)作为协变量，道德惩罚欲作为因变量进行协方差分析，结果显示，性别：$F(1, 204) = 0.34$, $p = 0.559$；年龄：$F(1, 204) = 1.13$, $p = 0.289$；民族：$F(1, 204) = 1.72$, $p = 0.191$；效应均不显著。组别的效应依然显著，$F(2, 204) = 9.59$, $p < 0.001$, $\eta_p^2 = 0.09$。

5.3　讨论

实验 4 通过直接操纵人们对算法的自由意志信念，将人类组与相信算法有自由意志组以及算法组进行比较，发现人类组与相信算法有自由意志组之间的差异，相较于人类组与算法组的差异有所减小，

从而进一步验证了自由意志信念是造成人们对不同歧视主体(人类 vs. 算法)产生不同道德惩罚欲的机制。那么，既然算法歧视比人类歧视引发较少道德惩罚欲的原因是人们认为算法比人类拥有较少的自由意志，那么拟人化倾向的个体差异是否会对这种效应起到调节作用呢？为了回答这一问题，在实验 5 中，我们将继续探索歧视行为主体影响道德惩罚欲的边界条件，考察拟人化倾向可能存在的调节效应。

6 实验 5：拟人化倾向的调节作用

理论上人类相较于算法有更高的自由意志，那么拟人化倾向较高的人是否会更倾向于增加对于算法自由意志的归因，从而对歧视行为主体对道德惩罚欲的关系造成影响呢？实验 5 旨在回答这一问题。此外，在实验 5 中我们着眼于年龄歧视问题，从而也进一步丰富了研究的歧视类型。

6.1 方法

6.1.1 被试

基于实验 5 适用的独立样本 *t* 检验，使用 G*Power 3.1 软件(Faul et al., 2007)计算本实验所需样本量，在显著性水平 α = 0.05 且中等效应量(*d* = 0.5)时，预测达到 90%统计检验力水平总共至少需要 172 名被试。通过 Credamo 平台招募被试，随机分配至人类组和算法组，实时剔除没有通过注意检查的被试数据并滚动采集，剩余有效数据共 199 名(女性 88 名)被试年龄在 18~41 岁(*M* = 28.64, *SD* = 4.68)之间，其中人类组 101 人，算法组 98 人。所有被试在实验开始之前均仔细阅读了实验说明并知情同意，有效数据被试在实验结束后获得一定实验报酬。

6.1.2 实验设计与程序

实验 5 为单因素两水平被试间实验设计，两组分别为人类组和算法组，所有被试被随机分配到其中一组。被试首先阅读年龄歧视的情境材料(下划线内容为人类组材料，括号中为算法组材料)：“在去年的秋季招聘中，韦蓝公司的人力资源经理赵广负责(使用算法)进行招聘。招聘结束后，公司发现赵广(算法)对应聘者的简历进行筛选时存在年龄偏差，筛掉了所有年龄大于 35 岁的应聘者，而公司的大部分岗位对年龄并无硬性要求。这阻碍了许多有才华、有能力、但年龄大于 35 岁的应聘者获得该公司的工作”，情境材料改编自 Bigman 等人(2020)的研究。

在阅读完情境材料并进行注意检查后，两组被试分别报告了对人类歧视或算法歧视的道德惩罚欲，测量条目同实验 1 (Hofmann et al., 2018), α = 0.84。然后，我们测量了被试对于情境中做出年龄歧视行为的人类或算法的自由意志信念，测量条目同实验 2 (Nadelhoffer et al., 2014), α = 0.87。接着，被试填写了拟人化个体差异量表(Individual Differences in Anthropomorphism Questionnaire, IDAQ; Waytz, Cacioppo, & Epley, 2014)，共 15 个条目(α = 0.87)，例如“普通的鱼在多大程度上有自由意志？”，采用李克特 11 点计分(从“0 = 一点也不”到“10 = 非常”)，得分越高表明被试的拟人化倾向越强。最后，被试报告了性别和年龄两项人口统计学信息。

6.2 结果

6.2.1 歧视行为主体对道德惩罚欲的影响

独立样本 *t* 检验结果显示，人类组的道德惩罚欲评分(*M* = 5.29, *SD* = 0.97)显著高于算法组(*M* = 4.61, *SD* = 1.32), *t*(197) = 4.17, *p* < 0.001, Cohen’s *d* = 0.59。为了进一步验证结果的稳健性，将被试的性别和年龄作为控制变量，方差分析结果显示，算法组的道德惩罚欲评分仍然显著低于人类组，*F*(1, 195) = 17.28, *p* < 0.001, η_p^2 = 0.08。

6.2.2 自由意志信念的中介效应

为了再次验证歧视主体对道德惩罚欲影响的心理机制，我们使用 Hayes (2013)提供的 SPSS 插件 PROCESS (Model 4)，以歧视主体为自变量(人类组 = −1，算法组 = 1)，自由意志信念为中介变量，道德惩罚欲为因变量，设定 Bootstrap 样本量为 5000，采用偏差校正的方法，选取 95%置信区间进行中介效应检验。数据结果显示，自由意志信念的中介效应值为−0.11，95%的 Bootstrap 置信区间为[−0.23, −0.01]，不包含 0，表明中介作用显著；并且在控制中介变量后，歧视主体对道德惩罚欲的直接效应为−0.23, 95%的 *Bootstrap* 置信区间为[−0.42, −0.04]，不包含 0，表明其直接效应仍然显著，自由意志信念在歧视主体对道德惩罚欲的影响中起部分中介作用。

6.2.3 拟人化倾向的调节效应

以道德惩罚欲为因变量考察歧视行为主体(人类组 = −1，算法组 = 1)与拟人化倾向的交互作用，结果表明歧视行为主体和拟人化倾向对道德惩罚欲存在显著的交互作用(*b* = 0.16, *SE* = 0.06, *t* = 2.70, *p* = 0.008)，人类组的道德惩罚欲显著高于算法组(*b* = −0.34, *SE* = 0.08, *t* = −4.18, *p* < 0.001)，拟人化倾向的高低对道德惩罚欲无显著影响(*b* = 0.01, *SE* =

0.06, t = 0.08, p = 0.937), 模型的调整 R^2 = 0.10, ΔR^2= 0.03, F (3, 195) = 8.40, p < 0.001。交互作用如图 4 所示, 简单斜率分析结果表明, 在低拟人化倾向条件下, 歧视行为主体对道德惩罚欲的影响显著(b = −0.57, SE = 0.12, t = −4.82, p < 0.001); 而在高拟人化倾向条件下, 歧视行为主体对道德惩罚欲的影响不显著(b = −0.12, SE = 0.12, t = −1.05, p = 0.295)。并且在算法组, 被试的拟人化倾向与对算法的自由意志信念显著正相关, r = 0.20, p = 0.044。

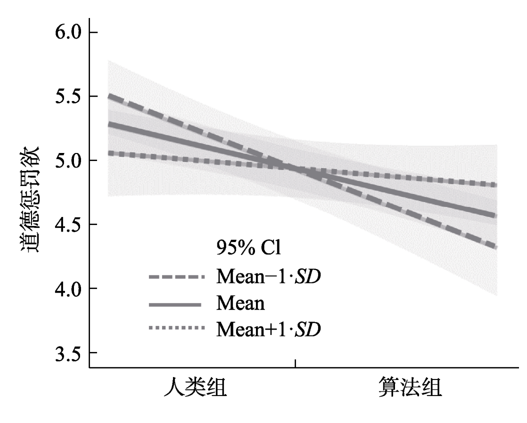


图 4　拟人化倾向的调节作用

6.3　讨论

实验 5 进一步探索了歧视行为主体影响道德惩罚欲的边界条件, 结果发现拟人化倾向在歧视行为主体对道德惩罚欲的影响中起调节作用。在拟人化倾向较低的被试中, 算法组的道德惩罚欲显著低于人类组; 而在拟人化倾向较高的被试中, 算法组和人类组的道德惩罚欲无显著差异。同时, 实验 5 还再次验证了自由意志信念的中介作用, 即人们之所以对算法歧视的道德惩罚欲较低, 是由于认为其与人类相比有更少的自由意志。在实验 6 中, 我们将直接操纵算法的拟人化程度, 以期进一步验证拟人化的调节作用。

7　实验 6：算法拟人化的调节作用

为了更直接地探讨算法拟人化是否会调节歧视主体对道德惩罚欲的影响, 实验 6 通过文字操纵的方式将算法拟人化, 比较被试对人类歧视、拟人化算法歧视以及非拟人化算法歧视的道德惩罚欲, 以此进一步验证拟人化在歧视主体影响道德惩罚欲中的调节作用。

7.1　方法

7.1.1　被试

基于实验 6 的单因素三水平被试间设计, 使用 G*Power 3.1 软件(Faul et al., 2007)计算本实验所需样本量, 在显著性水平 α = 0.05 且中等效应量(f = 0.25)时, 预测达到 90%统计检验力水平总共至少需要 207 名被试。通过 Credamo 平台招募被试, 随机分配至人类组、拟人化算法组和非拟人化算法组, 实时剔除没有通过注意检查的被试数据并滚动采集, 剩余有效数据共 207 名(女性 127 名), 被试年龄在 19~59 岁(M = 29.53, SD = 6.62)之间, 其中人类组、拟人化算法组和非拟人化算法组均为 69 人。所有被试在实验开始之前均仔细阅读了实验说明并知情同意, 有效数据被试在实验结束后获得一定实验报酬。

7.1.2　实验设计与程序

实验 6 为单因素三水平被试间实验设计, 三组分别为人类组、拟人化算法组以及非拟人化算法组, 所有被试被随机分配到其中一组。与实验 4 相同, 为了更好地排除被试对算法的理解程度可能对实验结果造成的影响, 两个算法组在歧视情境描述之前均加入了对“算法”含义的解释及举例说明, 以确保被试在完成实验前理解算法的含义。然后, 被试阅读了性别歧视的情境材料(下划线内容为人类组材料, 括号中为拟人化算法组和非拟人化算法组材料): “在去年的秋季招聘中, 韦蓝公司的人力资源经理赵广(使用算法“奇智”/使用算法“R2000”)负责进行招聘。招聘结束后, 公司发现赵广(奇智/R2000)对应聘者的简历进行筛选时存在性别偏差, 对于男性有明显的偏好, 筛掉了许多女性应聘者, 而公司的大部分岗位对性别并无硬性要求。这阻碍了许多有才华、有能力的女性应聘者获得该公司的工作”, 情境材料改编自 Bigman 等人(2020)的研究。

作为对算法拟人化的操纵, 两组算法组的被试在阅读完性别歧视的情境材料后还阅读了关于材料中所用招聘算法的介绍。其中, 拟人化算法组的被试阅读了以下关于算法“奇智”的介绍:

算法“奇智”的自我介绍:

嗨! 我叫奇智, 我是一种新型的招聘算法。我分析了过去十年投递给公司的所有简历, 来学习如何找出最优秀的应聘者。我能够仔细地审查应聘者的简历与背景, 准确预测未来可以满足岗位需求的员工、适合企业文化的员工, 找出最优秀的应聘者, 帮助企业挑选最好的员工。

非拟人化算法组的被试则阅读了以下关于算法“R2000”的介绍:

算法“R2000”的介绍:

R2000 是一种新型的招聘算法。R2000 分析了过去十年投递给公司的所有简历，来学习如何找出最优秀的应聘者。R2000 能够仔细地审查应聘者的简历与背景，准确预测未来可以满足岗位需求的员工、适合企业文化的员工，找出最优秀的应聘者，帮助企业挑选最好的员工。

该操纵参考了已有研究对于拟人化程度的操纵方式，即为非人对象起一个人名，并且以第一人称加以描述，能够有效地提升拟人化程度(如 Hur et al., 2015; May & Monga, 2014)。除此之外，两组对于算法的描述完全相同。为了检验拟人化操纵的有效性，被试需要对情境中算法的拟人化程度进行评分(“算法‘奇智’/‘R2000’在多大程度上让你想起了一些人类的特质？”，从“1 = 一点也没有”到“7 = 非常多”)，该操纵检查改编自 Hur 等人(2015)的研究。

在阅读完以上材料并进行操纵和注意检查后，三组被试分别报告了对人类歧视或算法歧视的道德惩罚欲，测量条目同实验 4(Hofmann et al., 2018)，评价对象为赵广/奇智/R2000, α = 0.88。最后，被试报告了性别和年龄两项人口统计学信息。

7.2 结果

7.2.1 拟人化操纵检查

独立样本 t 检验结果显示，拟人化算法组的拟人化评分(M = 5.43, SD = 0.88)显著高于非拟人化算法组(M = 4.83, SD = 1.25), $t(136)$ = 3.31, p = 0.001, Cohen’s d = 0.56。说明我们对算法拟人化的操纵有效。

7.2.2 算法拟人化的影响

以道德惩罚欲作为因变量进行单因素方差分析发现，歧视主体的主效应显著，$F(2, 204)$ = 12.60, $p < 0.001$, η_p^2 = 0.11。计划对比(planned contrast)分析表明，人类组的道德惩罚欲评分(M = 5.52, SD = 1.19, 95% CI [5.24, 5.81])显著高于拟人化算法组(M = 4.97, SD = 1.27, 95% CI [4.66, 5.27])和非拟人化算法组(M = 4.43, SD = 1.35, 95% CI [4.11, 4.76])，而拟人化算法组的道德惩罚欲评分也显著高于非拟人化算法组，$ps < 0.05$ (如图 5)。这表明对于同样的性别歧视，人类组(相比于算法组)的被试会产生更强烈的道德惩罚欲；而通过将算法拟人化，结果发现拟人化算法组(相比于非拟人化算法组)的被试也会产生更强烈的道德惩罚欲，说明算法拟人化在歧视主体对道德惩罚欲的影响中具有一定的调节作用。

7.3 讨论

实验 6 在实验 5 的基础上直接操纵的算法的拟人化程度，从而再次验证了算法拟人化在歧视主体对道德惩罚欲影响中的调节作用。具体而言，将算法拟人化会显著提升被试对算法的道德惩罚欲，这与我们的预测相一致。但拟人化算法组和人类组被试的道德惩罚欲仍然存在显著差异，这可能是由于文字操纵的算法拟人化虽然提升了算法整体的拟人化程度，但拟人化的算法仍然与人类水平有一定程度的差距(尤其是在自由意志信念方面)所致。

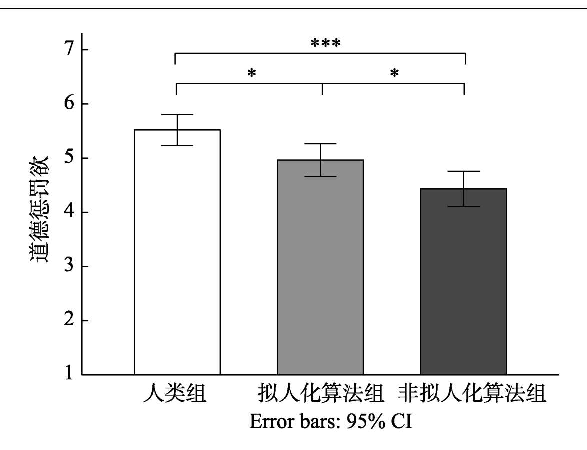


图 5 不同歧视主体组的道德惩罚欲评分

注：$^{*}p < 0.05$, $^{***}p < 0.001$.

8 总讨论

本研究考察了人们对人类歧视和算法歧视的道德惩罚欲是否存在差异，并在此基础上探讨了造成差异的潜在机制和边界条件。通过 6 项实验，我们发现相对于人类歧视，算法歧视会引发人们更少的道德惩罚欲，自由意志信念是造成道德惩罚欲差异的潜在机制，并且这一差异受到拟人化倾向的调节。具体而言，通过为不同被试呈现人类或算法同样的歧视行为并测量其道德惩罚欲，我们发现人们对算法歧视的道德惩罚欲小于对人类歧视的道德惩罚欲，并且这一差异具有稳健性(实验 1~6)。通过测量被试对人类或算法的自由意志信念(实验 2)以及对被试的自由意志信念(实验 3)和关于算法的自由意志信念(实验 4)进行操纵，我们进一步发现自由意志信念是造成人们对不同歧视主体(人 vs. 算法)产生不同道德惩罚欲的潜在机制，即人们认为算法与人类相比具有较少的自由意志，因此对算法歧视的道德惩罚欲较少(实验 2~4)。通过对被试拟人化倾向的测量(实验 5)和对算法拟人化程度的操纵(实验 6)，我们也发现了歧视主体对道德惩罚欲的影响受到拟人化的调节。在个体的拟人化倾向方面，对于拟人化倾向较低的被试来说，他们对算法歧视的道德惩罚欲小于对人类歧视的道德惩罚欲，

而对于拟人化倾向较高的被试来说，他们对算法歧视和人类歧视的道德惩罚欲不存在显著差异(实验5)；在算法的拟人化程度方面，算法的拟人化程度越高，人们对算法歧视与对人类歧视的道德惩罚欲差异越小。在研究中我们考察了不同类型的歧视，包括性别歧视(实验 1、6)、学历歧视(实验 2)、民族歧视(实验 3、4)以及年龄歧视(实验 5)；并且研究的样本涵盖了不同被试，包括来自 Credamo 平台(实验 1、2、5、6)的全国范围内被试以及来自某高校的大学生被试(实验 3、4)。实验情境材料和被试的多样性保障了研究结果的稳健性。

8.1　对人–算法的反应差异

本研究发现，在面对人类和算法做出同样的歧视行为时，人们会产生不同的道德惩罚欲，即人们对算法歧视的道德惩罚欲小于对人类歧视的道德惩罚欲。首先，这一结果与 Bigman 等人(2020)的研究发现相一致。Bigman 等人(2020)比较了人们对人类歧视和算法歧视的道德义愤，结果发现相对于人类歧视，人们对算法歧视的道德义愤较少，这一研究主要从道德情绪的方面探讨了人们对人类和算法反应的不对称性(asymmetry)。而本文则从行为倾向的视角出发，再次验证了人们对人类和算法反应的不对称性。但需要强调的是，虽然同为对不同歧视主体(人 vs. 算法)的比较，但和 Bigman 等人(2020)的研究相比，本研究与之区别的独特创新之处主要在于：第一，因变量的差异，Bigman 等人(2020)主要探讨了道德情绪即道德义愤的不同，而本文的研究重点在于行为倾向即道德惩罚欲的不同。虽然道德义愤和道德惩罚具有一定程度的相关性，但不意味着两者完全等同，道德惩罚欲依然有其独特的研究价值。并且 Bigman 等人(2020)虽然在研究 5 中加入了与道德惩罚测量具有一定相似性的条目(如“有歧视行为的算法应该被弃用”“采用该算法的公司应该道歉”)，但一方面这些条目并非完全针对算法，另一方面其研究结果也并未发现两组之间的显著差异。第二，机制的差异，Bigman 等人(2020)的研究着重探讨了动机机制，而本文则重复验证了自由意志信念的机制，并且从一定意义上来说自由意志机制比动机机制更为基础，因为个体具有自由意志是判断其动机的必要条件(Laming, 2004)；第三，调节的差异，Bigman 等人(2020)的研究并未进行调节变量的探索，他们在文章的局限和未来方向部分强调了拟人化倾向可能会扮演的调节作用，而本文则通过两个研究，从个体的拟人化倾向和算法本身的拟人化两个方面重复验证了拟人化的调节作用。

其次，这一发现将道德惩罚欲的相关研究拓展至人工智能领域。先前关于道德惩罚欲的研究大多聚焦于人类，且对于影响因素的探索也局限于人类相关变量(如 Hofmann et al., 2018)。本研究则在当前人工智能发展的大背景下扩大了道德惩罚欲的研究范围，将人工智能作为歧视行为主体的可能性纳入考察范围，并发现了歧视行为主体(人类 vs.算法)也会对道德惩罚欲产生显著影响。

此外，在对于算法决策的态度研究方面，我们的发现在一定程度上提供了算法厌恶(algorithm aversion)反例的新证据。算法厌恶研究发现，人们对算法在心理上有一种不信任感(Meehl, 1954)，虽然算法在计算能力等诸多方面已经超越了人类、表现更优，但通常情况下人们更偏好人类决策而非算法决策(Dietvorst et al., 2015, 2018)。如相对于人类决策的错误，算法决策错误对于人们来说更难以接受(Prahl & van Swol, 2017)。尤其是当决策涉及道德时，人们更反对机器代替人类做出决策，因为认为机器缺少做道德决策的必要心智能力，即使机器决策结果积极，人们仍然反对(Bigman & Gray, 2018)。在本研究中，我们发现当算法和人类都做出同样的歧视行为时，人们对算法的道德惩罚欲更小，这与先前支持算法厌恶的研究结果并不一致，如对算法犯错态度更严苛(Prahl & van Swol, 2017)等。但需要注意的是，这一结果也并不能够完全推翻算法厌恶的相关发现，因为人们或许仍然不愿意让算法做道德决策(Bigman & Gray, 2018)，但只是在算法和人类同样做出道德决策后消极反应更少而已。换句话说，人们对算法做道德决策的“欣赏”或许仅限于算法已经做出决策之后，而非体现在决策之前的偏好上。

要产生算法欣赏(algorithm appreciation)，似乎需要客观性较强的任务，人们才会更加偏好算法决策(Logg et al., 2019)。算法因其计算能力和客观性可能会被认为更准确公正(Grove et al., 2000)，但实际上算法歧视的事例并不鲜见，危险仍然存在(如 Borgesius, 2018)。在面对算法歧视时，人们不会像面对人类歧视时那般愤怒(Bigman et al., 2020)，我们也发现人们不会像面对人类歧视时那样想要进行惩罚，在情绪和行为倾向上程度都更低，由此可能造成对算法歧视的警惕性降低、习惯性甚至合理化增强，从而带来更严重的歧视问题。

8.2 对人−算法的知觉差异

本研究还发现人们对人类和算法的知觉存在差异，即认为算法比人类拥有更少的自由意志，而这也是解释道德惩罚欲差异的潜在机制。首先，这一发现与前人关于人工智能心智知觉和心理状态的相关研究大致一致(如 Gray et al., 2007)。如人们认为机器人具有中等程度的能动性(Epley & Waytz, 2010)，即做出自主、有计划行为的心理能力逊于人类。这与我们关于人们认为算法的自由意志少于人类的发现类似，都证实了人们认为算法等人工智能相关的机器型非人对象虽然有一定程度的心理能力，但仍远达不到人类水平。其次，这一发现也再次验证了自由意志信念与道德惩罚之间的紧密联系。人们之所以对算法歧视的道德惩罚欲小于对人类歧视的道德惩罚欲，我们发现与算法具有较低程度的自由意志有关。一个人自由行动、能够做出不同选择的能力对其是否应当承担道德责任和惩罚具有至关重要的影响(如 Shariff et al., 2014; Clark et al., 2014)，关于自由意志是否存在的信念也会影响对违规者的惩罚(Aspinwall et al., 2012)。本文的发现与前人对于自由意志和道德惩罚关联的研究结果一致，在更广义上，与心理能力是承担道德责任的必要前提这一观点也相符合(Gray et al., 2012)。此外需要提到的是，虽然可能存在与自由意志无关的竞争假设，如人的行为相对于算法更容易被解释、人无法真正惩罚算法等，但本文通过 2 个自由意志信念的调节效应研究(实验 3~4)和 2 个拟人化(与自由意志密切相关)的调节效应研究(实验 5~6)重复验证了本文所提出的自由意志信念机制，在一定程度上排除了上述竞争假设。

当然，对人−算法的知觉差异不仅体现在对其本身心智知觉的差异上，还可能体现在对其行为的知觉中。具体而言，虽然在实验情境甚至现实生活中人类和算法做出了完全一样的歧视行为并造成了同等影响，但人们对其严重性的知觉却可能存在差异。Bonezzi 和 Ostinelli (2021)的研究发现与人类歧视相比，做出性别和种族歧视行为的算法被认为有偏见的可能性更低，这是由于人们认为算法并不会像人类一样关注个体特征，而是以规则和程序进行笼统评判决策(Bonezzi & Ostinelli, 2021)。本文虽然没有直接测量和探讨人们对算法歧视和人类歧视严重性的知觉，但从研究的结果来看与已有发现是一致的。

同时，我们的研究发现拟人化倾向对于人们是否想要对算法歧视进行道德惩罚具有调节作用，这在本质上与具有完整的人类心智是承担道德责任的必要条件(Bigman & Gray, 2018; Gray et al., 2012)等研究结果相一致。人工智能的拟人化是业界趋势(Broadbent, 2017)、人工智能伦理问题之关键(Bostrom & Yudkowsky, 2011)。以歧视研究为例，以拟人化的算法作为歧视的主体不仅有助于更直接地探讨人们对人类和算法的知觉差异，或许也能够在一定程度上“模拟”人类歧视的心理过程，从而帮助我们更好地理解歧视的产生。

8.3 局限与展望

我们的研究证明了人们对人类歧视和算法歧视的道德惩罚欲存在差异，这种差异是由于人们对人类和算法的自由意志知觉不同，并且这一差异受到拟人化的调节。不过，我们也承认当前研究仍具有一定的局限性，这为未来的研究指出了一些方向。

首先，在实验设计的细节方面存在一定不足。第一，实验 1、2、3、5 的道德惩罚欲测量直接采用了 Hofmann 等人(2018)的条目，其中“道德惩罚”和“不道德行为”的表述置于算法歧视行为可能会对被试的回答造成一定的影响。当然我们在实验 4 和实验 6 中变换为了更加中性的说法，并未影响主要结果。第二，被试对算法的熟悉和了解程度可能对研究结果造成影响，但在一些实验(1、2、3、5)中我们并未解释算法的含义并举例说明，这在实验 4 和实验 6 中进行了弥补，但未来的研究仍需重点关注这一问题。第三，实验 2~4 的歧视情境描述存在歧视主体表述维度差异的问题，即“人力资源经理李原/张沛/赵广” 属于具体明确的对象，而算法属于宽泛的概念，无具体明确的对象。有研究发现，相比于一个身份不明的违规者，人们更倾向于惩罚一个确定的违规者(如 Small & Loewenstein, 2005)，所以实验中歧视主体表述维度的差异可能对研究结果造成一定的影响。为此我们在实验 1 和实验 4 的情境材料中均采用较为抽象的表述，在实验 6 的情境材料中则均采用较为具体的表述，均得到了类似结果。

其次，人们对人类歧视和算法歧视的道德惩罚欲差异可能还存在别的解释机制。在本文中我们着重探讨了自由意志信念的中介作用，但实际上人们对人类和算法心智知觉的差异还可能体现在别的方面，例如意识(consciousness; McDermott, 2007)、意向性(intentionality; Weisman et al., 2017)，以及体验情绪的心理能力(Epley & Waytz, 2010)等。未

来的研究可以更加细致地去考察这些可能的变量，并对这些影响因素进行比较，以便更透彻地理解人对人类和算法反应差异的心理机制。除了心智知觉方面的差异之外，上文中也提到人们对算法歧视和人类歧视的严重性程度的知觉存在差异(Bonezzi & Ostinelli, 2021)，因此在之后的研究中也可以将此纳入对人类和算法反应差异的机制探索中。此外，人们日常生活中约定俗成的“道德惩罚阈限”也可能对道德惩罚欲产生影响。即对于非人对象而言，人们对其道德惩罚欲可能本身就存在一定的阈限，哪怕其道德违规再严重也难以突破一定阈值，因此在之后的研究中也可以对相关阈限问题进行进一步的探讨。

最后，道德惩罚欲可能与道德归责的主体相关。在本文的研究中，我们考察了被试对算法的道德惩罚欲，发现人们对算法歧视的道德惩罚欲小于对人类歧视的道德惩罚欲。这一发现还有一个可能的原因在于人们并不愿意将道德责任归因于算法。以往研究发现人们不愿意让机器做道德决策(Bigman & Gray, 2018)，那么在机器已经做出道德决策的情况下，相应的道德责任究竟应该由谁来承担？是算法？是算法的设计者？是投资算法的企业？抑或监管算法的相关机构也需要承担一定的责任？随着人工智能的应用和算法决策的日益普及，越来越多的算法决策已经成为了既定事实，尽管人类仍然是这些人工智能的设计者、应用者、监管者，但伴随着人们对人工智能道德谴责的增加，人工智能相应承担的责任和惩罚也会加重，这无疑会带来一种潜在的“甩锅”可能性——设计者、企业甚至政府利用人工智能来逃避自身错误的责任(Bigman et al., 2019)。而道德归责之后的惩罚也会随之受到影响，这一方面是由于责任与惩罚的相关性，即人们若倾向于将责任归因于人，那么对于人工智能的道德惩罚也会相应较少；另一方面，从道德惩罚本身的“惩戒”、“规范”等实际功用而言，对人工智能背后的“人”，尤其是涉及非法牟利或使用不当的“使用者”进行惩罚或许更合理，也更有现实意义。虽然如算法等人工智能现在还无法成为完全的道德主体，但是在其犯错时，我们依然还是可能惩罚它，如扫地机器人犯错可能招致人类脚踢的惩罚，而若算法犯错，人也可能惩罚承载其的智能设备，如摔手机等。当然，对算法探讨的核心也是为了人类福祉。因此，对于人类和人工智能之间道德责任乃至道德惩罚的分配都值得进一步研究。

9　结论

本研究结论如下：第一，相对于人类歧视，人们对算法歧视的道德惩罚欲更少；第二，这一现象的潜在机制是人们认为算法(与人类相比)更缺乏自由意志；第三，个体拟人化倾向越强或者算法越拟人化，人们对算法的道德惩罚欲越强。

参　考　文　献

Al Ramiah, A., Hewstone, M., Dovidio, J. F., & Penner, L. A. (2010). The social psychology of discrimination: Theory, measurement and consequences. In L. Bond, F. McGinnity, & H. Russell (Eds.), *Making equality count: Irish and international research measuring equality and discrimination* (pp. 84−112). Dublin, Ireland: Liffety Press.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias*. Retrieved June 18, 2021, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Angwin, J., Mattu, S., & Larson, J. (2015). *The tiger mom tax: Asians are nearly twice as likely to get a higher price from Princeton Review*. Retrieved June 18, 2021, from https://www.ProPublica.org/article/asians-nearly-twice-as-likely-to-get-higher-price-from-princeton-review

Aspinwall, L. G., Brown, T. R., & Tabery, J. (2012). The double-edged sword: Does biomechanism increase or decrease judges' sentencing of psychopaths? *Science, 337*(6096), 846−849.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*(1), 71−81.

Batson, C. D., Kennedy, C. L., Nord, L.-A., Stocks, E. L., Fleming, D. Y. A., Marzette, C. M., ... Zerger, T. (2007). Anger at unfairness: Is it moral outrage? *European Journal of Social Psychology*, *37*(6), 1272−1285.

Baumeister, R. F. (2008). Free will in scientific psychology. *Perspectives on Psychological Science*, *3*(1), 14−19.

Baumeister, R. F. (2014). Constructing a scientific theory of free will. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Vol. 4. Free will and moral responsibility* (pp. 235−255). Boston Review.

Baumeister, R. F., Masicampo, E. J., & DeWall, C. N. (2009). Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. *Personality and Social Psychology Bulletin*, *35*(2), 260−268.

Baumeister, R. F., Stillwell, A., & Wotman, S. R. (1990). Victim and perpetrator accounts of interpersonal conflict: Autobiographical narratives about anger. *Journal of Personality and Social Psychology*, *59*(5), 994−1005.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181,* 21−34.

Bigman, Y. E., Gray, K., Waytz, A., Arnestad, M., & Wilson, D. (2020). *Algorithmic discrimination causes less moral outrage than human discrimination* [Preprint]. PsyArXiv.

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences*, *23*(5), 365−368.

Bonezzi, A., & Ostinelli, M. (2021). Can algorithms legitimize

discrimination? *Journal of Experimental Psychology: Applied*, *27*(2), 447−459.

Borgesius, F. Z. (2018). Discrimination, artificial intelligence, and algorithmic decision-making. Retrieved June 18, 2021, from https://rm.coe.int/discrimination-artificial-intelligence-andalgorithmic-decision-making/1680925d73

Bostrom, N., & Yudkowsky, E. (2011). The ethics of Artificial Intelligence. In K. Frankish (Ed.). *Cambridge handbook of artificial intelligence*. Cambridge: Cambridge University Press.

Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. *Annual Review of Psychology*, *68*, 627−652.

Clark, C. J., Luguri, J. B., Ditto, P. H., Knobe, J., Shariff, A. F., & Baumeister, R. F. (2014). Free to punish: A motivated account of free will belief. *Journal of Personality and Social Psychology*, *106*(4), 501−513.

Correll, J., Judd, C. M., Park, B., & Wittenbrink, B. (2010). Measuring prejudice, stereotypes and discrimination. *The SAGE handbook of prejudice, stereotyping and discrimination*, (pp. 45−62). Thousand Oaks, CA: Sage.

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. Retrieved June 18, 2021, from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies, 1*(1), 92−112.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114−126.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155−1170.

Epley, N., & Waytz, A. (2010). Mind perception. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (pp. 498−541). John Wiley & Sons, Inc.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864−886.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175−191.

Ferrero, F., & Barujel, A. G. (2019, October). Algorithmic driven decision-making systems in education: Analyzing bias from the sociocultural perspective. In *2019 XIV Latin American Conference on Learning Technologies (LACLO)* (pp. 166−173), San Jose Del Cabo, Mexico.

Freedman, R., Borg, J. S., Sinnott-Armstrong, W., Dickerson, J. P., & Conitzer, V. (2020). Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence*, *283*, 103261.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*(2), 101−124.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19−30.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20*, 98−116.

Hao, K. (2019). *AI is sending people to jail—and getting it wrong.* Retrieved June 18, 2021, from https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/

Harvey, C. R., Rattray, S., Sinclair, A., & van Hemert, O. (2017). Man vs. machine: Comparing discretionary and systematic hedge fund performance. *The Journal of Portfolio Management, 43*(4), 55−69.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York: Guilford Press.

Hofmann, W., Brandt, M. J., Wisneski, D. C., Rockenbach, B., & Skitka, L. J. (2018). Moral punishment in everyday life. *Personality and Social Psychology Bulletin*, *44*(12), 1697−1711.

Hur, J. D., Koo, M., & Hofmann, M. (2015). When temptations come alive: How anthropomorphism undermines self-control. *Journal of Consumer Research, 42*(2), 340−358.

Kim, T. W., & Duhachek, A. (2020). Artificial intelligence and persuasion: A construal-level account. *Psychological Science, 31*(4), 363−380.

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, *65*(7), 2966−2981.

Laming, D. (2004). *Understanding human motivation: What makes people tick?* Malden, MA: Blackwell.

Leo, X., & Huh, Y. E. (2020). Who gets the blame for service failures? Attribution of responsibility toward robot versus human service providers and service firms. *Computers in Human Behavior, 113*(4), 106520.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90−103.

Mackenzie, M. J., Vohs, K. D., & Baumeister, R. F. (2014). You didn't have to do that: Belief in free will promotes gratitude. *Personality and Social Psychology Bulletin, 40*(11), 1423−1434.

May, F., & Monga, A. (2014). When time has a will of its own, the powerless don't have the will to wait: Anthropomorphism of time can decrease patience. *Journal of Consumer Research, 40*(5), 924−942.

McDermott, D. (2007). Artificial intelligence and consciousness. In P. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 117−150). Cambridge: Cambridge University Press.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Mori, M. (1970). The uncanny valley. *Energy, 7,* 33−35.

Nadelhoffer, T., Shepard, J., Nahmias, E., Sripada, C., & Ross, L. T. (2014). The free will inventory: Measuring beliefs about agency and responsibility. *Consciousness and Cognition*, *25*, 27−41.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology, 18*(1), 561−584.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs, 41*(4), 663−685.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to

manage the health of populations. *Science, 366*(6464), 447−453.
Prahl, A., & van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, *36*(6), 691−702.
Rigoni, D., Kühn, S., Gaudino, G., Sartori, G., & Brass, M. (2012). Reducing self-control by weakening belief in free will. *Consciousness and Cognition*, *21*(3), 1482−1490.
Shariff, A. F., Greene, J. D., Karremans, J. C., Luguri, J. B., Clark, C. J., Schooler, J. W., ... Vohs, K. D. (2014). Free will and punishment: A mechanistic view of human nature reduces retribution. *Psychological Science*, *25*(8), 1563−1570.
Sinnott-Armstrong, W. (2014). *Moral psychology: Free will and moral responsibility.* MIT Press.
Small, D. A., & Loewenstein, G. (2005). The devil you know: The effects of identifiability on punishment. *Journal of Behavioral Decision Making, 18*(5), 311−318.
Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science*, *19*(1), 49−54.
Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219−232.
Waytz, A., Cacioppo, J., & Epley, N. (2014). Who sees human?: The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*(3), 219−232.
Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113−117.
Wegner, D. M., & Gray, K. (2017). *The mind club: Who thinks, what feels, and why it matters*. Penguin.
Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people’s conceptions of mental life. *Proceedings of the National Academy of Sciences*, *114*(43), 11374−11379.
Wen, Z., Zhang, L., Hou, J., & Liu, H. (2004). Testing and application of the mediating effects. *Acta Psychologica Sinica, 36*(5), 614−620.
[温忠麟, 张雷, 侯杰泰, 刘红云. (2004). 中介效应检验程序及其应用. *心理学报, 36*(5), 614−620.]

Algorithmic discrimination causes less desire for moral punishment than human discrimination

XU Liying[1], YU Feng[2], PENG Kaiping[3]

([1] *School of Marxism, Tsinghua University, Beijing* 100084, *China*) ([2] *Department of Psychology, School of Philosophy, Wuhan University, Wuhan* 430072, *China*) ([3] *Department of Psychology, School of Social Sciences, Tsinghua University, Beijing* 100084, *China*)

Abstract

The application of algorithms is believed to contribute to reducing discrimination in human decision-making, but algorithmic discrimination still exists in real life. So is there a difference between folk responses to human discrimination and algorithmic discrimination? Previous research has found that people's moral outrage at algorithmic discrimination is less than that at human discrimination. Few studies, however, have investigated people's behavioral tendency towards algorithmic discrimination and human discrimination, especially whether there is a difference in their desire for moral punishment. Therefore, the present study aimed at comparing people's desire to punish algorithmic discrimination and human discrimination as well as finding the underlying mechanism and boundary conditions behind the possible difference.

To achieve the research objectives, six experiments were conducted, which involved various kinds of discrimination in daily life, including gender discrimination, educational background discrimination, ethnic discrimination and age discrimination. In experiment 1 and 2, participants were randomly assigned to two conditions (discrimination: algorithm vs. human), and their desire for moral punishment was measured. Additionally, the mediating role of free will belief was tested in experiment 2. To demonstrate the robustness of our findings, the underlying mechanism (i.e., free will belief) was further examined in experiment 3 and 4. Experiment 3 was a 2 (discrimination: algorithm vs. human) × 2 (belief in free will: high vs. low) between-subject design, and experiment 4 was a single-factor (discrimination: human vs. algorithm with free will vs. algorithm without free will) between-subject design. Experiment 5 and 6 were conducted to test the moderating role of anthropomorphism. Specifically, participants’ tendency to anthropomorphize was measured in experiment 5, and the anthropomorphism of algorithm was manipulated in experiment 6.

As predicted, the present research found that compared with human discrimination, people have less desire to punish algorithmic discrimination. And the robustness of this result was demonstrated by the diversity of our

stimuli and samples. In addition, we found that free will belief played a mediating role in the effect of discrimination (algorithm vs. human) on the desire to punish. That is to say, the reason why people had less desire to punish when facing algorithm discrimination was that they thought algorithms had less free will than humans. Finally, the results also demonstrated the moderating effect of anthropomorphism.

These results enrich literature regarding algorithm discrimination as well as moral punishment from the perspective of social psychology. First, this research explored people's behavioral tendency towards algorithmic discrimination by focusing on the desire for moral punishment, which contributes to a better understanding of people's responses to algorithmic discrimination. Second, the results are consistent with previous studies on people’s mind perception of artificial intelligence. Third, it adds evidence that free will has a significant impact on moral punishment.

Key words algorithm, algorithmic discrimination, moral punishment, free will belief, anthropomorphism